When some triggers a scalar inference out of the blue

Barbet, Cécile; Thierry, Guillaume

Cognition

DOI:
10.1016/j.cognition.2018.03.013

Published: 01/08/2018

Peer reviewed version

When *some* triggers a scalar inference out of the blue An electrophysiological study of a Stroop-like conflict elicited by single words

Cécile Barbet[a,*], Guillaume Thierry[b]

[a] *Université de Genève, Faculté de Psychologie et des Sciences de l'Education, Uni Pignon, 42 Boulevard du Pont d'Arve, CH-1205 Genève*

[b] *Bangor University, School of Psychology, Brigantia Building, LL57 2AS Bangor, UK*

Abstract

Some studies in experimental pragmatics have concluded that scalar inferences (e.g., '*some* X are Y' implicates '*not all* X are Y') are context-dependent pragmatic computations delayed relative to semantic computations. However, it remains unclear whether strong contextual support is necessary to trigger such inferences. Here we tested if the scalar inference 'not all' triggered by *some* can be evoked in a maximally neutral context. We investigated event-related potential (ERP) amplitude modulations elicited by Stroop-like conflicts in participants instructed to indicate whether strings of letters were printed with all their letters in upper case or otherwise. In a randomized stream of non-words and distractor words, the words *all*, *some* and *case* were either presented in capitals or they featured at least one lower case letter. As expected, we found a significant conflict-related N450 modulation when comparing e.g., 'aLl' with 'ALL'. Surprisingly, despite the fact that most responses from the same participants in a sentence-picture verification task were literal, we also found a similar modulation when comparing 'SOME' with e.g., 'SoMe', even though SOME could only elicit such a Stroop conflict when construed pragmatically. No such modulation was found for e.g., 'CasE' vs. 'CASE' (neutral contrast). These results suggest that *some* can appear incongruent with the concept of

*Corresponding author
Email address: `Cecile.Barbet@unige.ch` (Cécile Barbet)

'all' even when contextual support is minimal. Furthermore, there was no significant correlation between N450 effect magnitude ('SOME' minus e.g., 'sOMe') and pragmatic response rates recorded in the sentence-picture verification task. Overall, this study shows for the first time that the pragmatic meaning of *some* can be accessed in a maximally neutral context, and thus, that the scalar inference 'not all' triggered by *some* should be construed as context-*sensitive* rather than context-*dependent*, that is, more or less salient and relevant depending on the context rather than entirely contingent upon it.

Keywords: Experimental semantics and pragmatics, non-literal meaning, context-dependency, Stroop, event-related brain potentials, N450 effect

---

1. Introduction

Consider the following exchange:

(1) A: What time is it?
B: Some of the guests are already leaving. (Levinson, 2000, p. 16)

From B's answer, it can be expected that A will understand that (i) it must be late, and (ii) not all of the guests are already leaving (see Levinson, 2000, pp. 16-17). Both (i) and (ii) contribute to the pragmatic rather than literal meaning of B's utterance and are called *implicatures* (see e.g., Grice, 1975; Levinson, 2000; Sperber and Wilson, 1995). However, implicatures like (ii) can be derived because of the mere presence of particular words such as *some*, whereas implicatures like (i) require a specific context and can only be derived from the complete utterance. The difference is made apparent when changing A's question into "Where is John?" for example (see e.g., Levinson, 2000, p. 17), in which case implicature (ii) remains valid, whereas implicature (i) does not.

In Grice's (1975) terms, implicatures such as (i) are Particularized Conversational Implicatures (henceforth PCIs) and those such as (ii) are Generalized Conversational Implicatures (GCIs). A particular case of GCI is the *scalar implicature* or *scalar inference* (hereafter SI), which is triggered when a linguistic expression has a stronger competitor along a scale of informativeness (see e.g., Horn, 1972; Horn, 1989; Levinson, 2000). For instance, in (1), *some* contrasts with *all* and thus can trigger the SI 'not all'. Other examples of such lexical scales are ⟨always, sometimes⟩, ⟨necessarily, possibly⟩, ⟨and, or⟩, ⟨finish, start⟩, ⟨love, like⟩, ⟨hot, warm⟩ (see e.g., Levinson, 2000; van Tiel, van Miltenburg et al., 2014).

Following the footsteps of Grice, some scholars endorsed the GCI – PCI distinction and argued that a GCI is the preferred or standard interpretation of a word such as *some* "in the absence of special circumstances" (Grice, 1975, p. 56), relating to "a *default* mode of reasoning" (Levinson, 2000, p. 42). A GCI remains an implicature since it is *defeasible*, that is it can be cancelled without resulting in a contradiction, contrary to the case of literal meaning ('at least one' as for *some*):

(2) *Some* of the students failed. *In fact all* of them failed.
#*Some* of the students failed. *In fact none* of them failed.

Other scholars have argued that the distinction between GCI and PCI is unfounded, because all implicatures, including SIs, should be considered particularized (see notably Sperber and Wilson, 1995; Carston, 2004).

The two views presented above have been referred as 'default' and 'context-driven' models of SI derivation (see e.g., Breheny, Katsos and Williams, 2006; Politzer-Ahles and Fiorentino, 2013). The former view predicts that SIs are *relatively* context-independent, realised immediately and effortlessly. The latter view predicts that SIs are context-dependent, only realised in contexts in which they are relevant, derived after the processing of semantic meaning and context, and require additional cognitive effort.

Some studies in experimental pragmatics have concluded that SIs are cognitively costly (see e.g., Bott and Noveck, 2004; De Neys and Schaeken, 2007; Chevallier et al., 2008) and akin to context-dependent pragmatic computations (see e.g., Breheny, Katsos and Williams, 2006) that are delayed relative to semantic computations (see e.g., Huang and Snedeker, 2009). However, other studies have shown that SIs are not necessarily delayed (see e.g., Grodner et al., 2010; Degen and Tanenhaus, 2015), that the cognitive cost associated with them might not stem from the inferencing process itself (Marty and Chemla, 2013; Chemla and Bott, 2014), and that a strong contextual support may not be needed (Politzer-Ahles and Gwilliams, 2015).

In the present study, we focused on the dependency of SI derivation upon the context. In experimental pragmatics, SI context-dependency has been mostly investigated in reading (see e.g., Breheny, Katsos and Williams, 2006; Politzer-Ahles and Fiorentino, 2013) by comparing SI-supportive and SI-non-supportive contexts. In an SI-supportive context, the SI answers the 'question under discussion' (henceforth QUD, see e.g., Roberts, 1996; Beaver and Clark, 2008), whereas in an SI-non-supportive context, the SI does not answer the QUD. In, e.g.:

(3) Mary was preparing to throw a party for John's relatives. She asked John whether *all*/*any* of them were staying in his apartment. John said that *some of them* were. He added that *the rest* would be staying in a hotel. (Politzer-Ahles and Fiorentino, 2013)

the predictions drawn from the default and context-driven models are the following: in non-supportive contexts ('*any*'), the SI is not available, therefore '*the rest*' requires more processing time than in supportive contexts ('*all*') in which '*not all* of the relatives' becomes relevant and facilitates the bridging inference that *the rest* means 'the rest of the relatives'. The two models differ regarding their account of SI unavailability in non-supportive contexts: either it is not derived because it is not relevant in the context (context-driven model), or it is automatically derived and then cancelled once it becomes apparent that it is not relevant (default model). Therefore, the context-driven model predicts an increase in reading time at the segment containing *some* in SI-supportive contexts relative to SI-non-supportive contexts. Contrastingly, the default model predicts no difference between conditions since the SI should automatically be derived in both cases. In self-paced reading studies such as Breheny, Katsos and Williams (2006) or Politzer-Ahles and Fiorentino (2013), an increase in reading time for the anaphoric noun phrase (e.g., *the rest*) was recorded in SI-non-supportive contexts relative to SI-supportive contexts. However, the increase in processing for the *some*-region predicted by the context-driven model was observed in Breheny, Katsos and Williams' (2006) study (see also Bergen and Grodner, 2012), but not in Politzer-Ahles and Fiorentino's (2013) study (see also Hartshorne et al., 2015). In sum, the results obtained for the anaphoric noun phrase suggest that the SI is context-*sensitive* and thus more salient in supportive contexts in which it answers the QUD. However, it is remains unclear whether SI derivation is actually context-*dependent*.

At this point, we note that Levinson's (2000) default model of GCIs might have been inadequately interpreted in the literature. Levinson uses the expression "implicature cancellation" in the case of:

(4) A: "Is there any evidence against them?"
B: "Some of their identity documents are forgeries." (Levinson, 2000, p. 51)

However, this example of cancellation is given as an example of "predicted (but nonoccurring) scalar implicature", that is, an example of a case in which "we do not let the inference through. That's because, intuitively, A is only interested in whether there is at least some evidence against the criminals; given A's question, all that is relevant is the possession of at least some evidence" (Levinson, 2000, p. 51). Levinson concludes: "It seems then that Relevance implicatures, or inferences about the speaker's goals, can limit the amount of further inference that is warranted. Thus even where these further inferences are entirely consistent with all that is known, they do not go through." (Levinson, 2000, p. 52). In other words, it seems that Levinson makes a similar prediction as context-driven models: if the SI is irrelevant in the discourse context, it does not arise. Non-supportive *any*-contexts may thus be part of these "special circumstances" (see above) in which SIs are not derived.

In a recent study, Politzer-Ahles and Gwilliams (2015) asked participants whether it was possible, in an example such as (5), that all of John's relatives stay in his apartment:

(5) Mary was preparing to throw a party for John's relatives. She asked John whether *all*/*any* of them were staying in his apartment. John said that *some of them* were.

Only slightly more than 20 % of the responses were 'yes' (see Corrigendum), in either the *all*-contexts or the *any*-contexts. Thus, contrary to the predictions of both default and context-driven models, the SI 'not all' tends to be computed in *any*-contexts also, even though it is not relevant. Therefore, it appears that SI derivation does not in fact require much context support. Reciprocally, supposedly "blocking" contexts such as *any*-contexts do not guarantee that the SI will not be derived. MEG results from the same study (Politzer-Ahles and Gwilliams, 2015) using the same stimuli, showed greater activation for *some* in non-supportive contexts compared with supportive contexts, suggesting greater effort to derive the SI in non-supportive contexts. However, it is possible that the 'not all' interpretation may have been constrained by the presence of *of* following *some* in both contexts. Indeed, Grodner et al. (2010, Appendix A) showed that the partitive *some of* is more likely to give rise to the SI than the bare quantifier *some* (see also Geurts, 2010, p. 100; Degen and Tanenhaus, 2015).

Another way of investigating SI context-dependency would be to use *neutral* contexts, that is, contexts that are unbiased towards a lower or upper-bounded interpretation of *some*. This was the aim of Breheny, Katsos and Williams' (2006) second self-paced reading experiment. In this experiment, there was no preceding context to the sentence containing *some*, however '*some of the*' + noun was the grammatical subject or object of the sentence, that is, it was either in a topical or non-topical position, respectively. The sentences containing *some* were followed by sentences beginning with a noun phrase meaning ‘the rest’ or ‘the others’:

(6) The director had a meeting with *some of the consultants.*/*Some of the consultants* had a meeting with the director. *The rest* did not manage to attend.

If the SI ‘*not all* consultants met with the director’ is triggered by default, the referent of ‘*the rest*’ should be equally accessible no matter where ‘*some of the consultants*’ is located in the preceding sentence. In contrast, if the SI is context-dependent, the referent of ‘*the rest*’ should be more accessible when ‘*some of the consultants*’ is in a topical position, resulting in shorter reading times for the anaphoric noun phrase ‘*the rest*’. This predicted difference was observed. However, it can be argued (i) that more time was available to compute the SI when its trigger was at the beginning rather than the end of the sentence and therefore that the SI was more likely to influence the processing of the anaphoric segment in the former case than the latter, and (ii) that the contexts were not genuinely neutral, because utterances appear to elicit an implicit context when they are presented in isolation (see e.g., Geurts, 2010, p. 91, 98; Tian, Ferguson and Breheny, 2016). Therefore, when ‘*some of the consultants*’ is the subject of the preceding sentence, it is more likely that the topic of the implicit discourse context concerns the consultants rather than the director, and thus that it answers the QUD as compared to when it is in non-topical position. In other words, sentence reading without stated context is not entirely free of contextual effects.

In sum, reading studies have shown that SIs are context-*sensitive* (see also Bonnefon, Feeney and Villejoubert, 2009; Bergen and Grodner, 2012; Breheny, Ferguson and Katsos, 2013; Goodman and Stuhlmüller, 2013) as predicted by both default and context-driven models. However, context-*dependency* remains undemonstrated, probably owing to the fact that the context of an utterance can essentially never be neutral. In other words, the question addressed here is: do SIs need strong contextual support to be accessed? Thus, in the present study, we elected to present the quantifier *some* in isolation in order to test the context-dependency of the SI ‘not all’ when context is maximally neutral. It is indeed highly unlikely that words presented in isolation would be contextualised to any great extent. We asked participants to perform a task unrelated to the meaning of the words presented and investigated whether the ‘not all’ inference would hinder the expected response via Stroop-like interference in an event-related brain potentials (ERPs) experiment.

Most of previous ERP studies have investigated the time course of SIs (see e.g., Noveck and Posada, 2003; Nieuwland, Ditman and Kuperberg, 2010; Hunt et al., 2013), their processing cost (see e.g., Hartshorne et al., 2015), or their neural correlates (see e.g., Politzer-Ahles, Fiorentino et al., 2013), but, to our knowledge, no previous study has tested context-dependency directly, and *a fortiori* N450 effects in a Stroop-like paradigm.

In a typical colour-word Stroop task, participants are asked to name the physical colour of a word which can either be congruent with the meaning of the word (e.g., BLUE presented in blue) or incongruent (e.g., BLUE presented in red). Performance is usually worse, and naming latencies usually slower, for incongruent trials as compared with congruent (or neutral, e.g., SHOE presented in blue) trials. Such interference appears to arise from parallel analysis of task-relevant (physical colour) and task-irrelevant (word meaning) stimulus dimensions (for a review, see MacLeod, 1991). ERP studies investigating such interference have primarily identified a robust N450 effect (see e.g., West, 2003; Markela-Lerenc et al., 2004; Szűcs and Soltész, 2010; Tillman and Wiens, 2011), alternatively labelled Ni (negativity for incompatible Stroop trials, Eppinger et al., 2007), N400 (Rebai, Bernard and Lannou, 1997) or MFN (medial frontal negativity, Chen et al., 2011). This conflict-sensitive ERP modulation manifests as an increased negativity over fronto-central to centro-parietal regions between 300–500 ms after stimulus onset for incongruent as compared with congruent (or neutral) trials. This effect is found in a variety of contexts, e.g., overt verbal response, covert verbal response, or manual response (see e.g., Liotti et al., 2000), as well as in numerical Stroop tasks (see e.g., Szűcs, Soltész and White, 2009).

Here, we presented the words *all*, *some* and *case* in isolation amongst non-words and distractor words. The stimuli were presented either with all letters in upper case (e.g., ALL) or as a mix of upper and lower cases letters (e.g., aLl). Participants were asked to indicate whether or not each stimulus was presented with all its letters in upper case. Over and above a classical effect of physical form (mixed vs. upper case) expected from the literature (see e.g., Mayall, Humphreys and Olson, 1997; Mayall, Humphreys, Mechelli et al., 2001; Juhasz et al., 2006; Arditi and Cho, 2007; Lien, Allen and Crawford, 2012), in the case of stimuli such as 'alL', we expected greater conflict as compared to the case of 'ALL', manifesting as a modulation of the N450, because of the incongruence between the word's guise and its meaning. As regards *some*, two scenarios were possible. *Some* can either mean 'some and possibly all', or 'some but not all' if an SI is derived. On the one hand, if SI derivation is strongly context-dependent, the stimuli 'SOME' and 'SoMe', for instance, should yield similar responses because the meaning would not be incongruent with the expected response. On the other hand, if *some* can evoke 'not all' when contextual support is minimal, an N450 modulation should be expected for 'SOME' as compared to 'SoMe'. In the latter situation, the difference between 'SOME' and 'SoMe' should resemble the difference observed between 'aLl' and 'ALL'. Finally, we presented the stimulus 'CASE' and variants such as 'CaSe' as a neutral control with the expectation that no Stroop conflict would occur for this word.

We also investigated participant's pragmatic behaviour in a task similar to that used by van Tiel and Schaeken (2016) to measure literal vs. pragmatic response styles in the presence of contextual information. In this second experiment, participants had to indicate whether under-informative *some*-statements (e.g., "Some of the circles are green.") provided a good description of a situation depicted by means of a figure (e.g., a mixed line-up of circles and squares in which all circles are green). Results from this off-line experiment allowed us to compare spontaneous processing of *some* indexed by ERPs in a minimal context condition with pragmatic behaviour in a context-rich condition.

2. Methods

2.1. Participants

Thirty-four native speakers of English (20 females; mean age = 21.8, SD = 7.2) gave written consent to take part in the experiment approved by the Ethics Committee of Bangor University, United Kingdom. All were students from the School of Psychology and were given course credits for their participation. All had normal or corrected-to-normal vision. The data of 7 participants had to be dismissed due to excessive artefacts in the EEG recordings (see section 2.2.3 for details). Therefore, statistical analyses of behavioural and ERP results are based on 27 individual data sets.

2.2. ERP experiment

2.2.1. Materials

Critical stimuli were the words *some* (pragmatic test), *all* (semantic test), and *case* (neutral control) intermixed with three or four-letter filler non-words (rop, fusk, cauv, urbe, tarb, demb, soys, tovs, gyte, kilv) and four-letter distractor words (*font*, *zero*, *each*, *none*, *most*) so as to avoid directing special attention to the critical stimuli. All stimuli appeared in Courier New size 14 points subtending approximately 1 degree of visual angle at the centre of a 19” CRT monitor with all letters in upper case or as a mix of upper and lower case letters (see **Fig. 1**).

The analysis focused on two factors: *word-type* (*all*, *some*, *case*) and *case-type* (upper case, mixed case), resulting in a 3 × 2 experimental design. In the *mixed case* condition, the number and position of upper case letters systematically varied within words and non-words (i.e., aLl, alL, ALl, aLL, AlL; SOme, SoMe, sOMe, soME, sOmE; etc.). ‘All’, ‘Some’, ‘Case’, ‘Font’, etc. were avoided because of regular sentence case, and *some* with only one upper case letter was

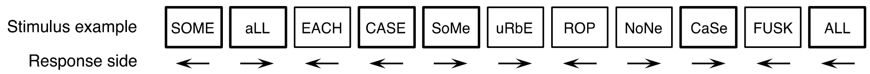

Figure 1: **Example of stimulus sequence and expected responses.** Critical test stimuli have a bold frame. Note that critical stimulus density is exaggerated in this example in order to show all six types of critical stimuli.

also avoided because of the ambiguity surrounding its minimal meaning ('at least one' or 'more than one'?). Importantly, stimulus congruence was manipulated in opposition to *case-type* for the two critical stimuli *all* and *some*, since 'ALL' was semantically congruent with its guise whereas 'SOME' was pragmatically incongruent with its guise.

2.2.2. Procedure

Participants were told that they would see strings of letters presented one at a time. They were asked to indicate as quickly and as accurately as possible whether or not all letters were capital letters by pressing designated left and right buttons on a response box. Response side was counterbalanced between participants. Stimuli were presented in a fully randomized sequence with the words *some*, *all* and *case* appearing 60 times in upper case (CAPS condition) and 60 times as a mix of upper and lower case letters (MiXEd condition). Each of the four-letter filler non-words appeared 22 times in each of the CAPS and MiXEd conditions, the three-letter filler non-words appeared 24 times and each of the 5 distractor words appeared 60 times, leading to a total of 1044 filler/distractor trials and 360 test trials. Each stimulus was displayed for 1000 ms or until participant's response, whichever was the shortest, with a randomly selected inter-stimulus interval of 460, 480, 500, 520, or 540 ms in order to reduce cross-trial ERP contamination. A training phase preceded the experiment involving 36 exemplars of the stimuli. Participants needed less than 30 minutes to complete the task in 4 blocks, and were invited to take a break between blocks at their convenience.

2.2.3. EEG recording and analysis

Electroencephalographic (EEG) data were recorded continuously at a rate of 1 kH in reference to electrode Cz with an online bandpass filter set between 0.01–200 Hz from 64 Ag/AgCl electrodes using SynAmp2 amplifiers (Neuroscan Inc., El Paso, TX, USA). Electrodes were attached to an elastic cap (Easycap™, Herrsching, Germany) and placed according to the extended 10-20 convention. The ground electrode was placed at FPz. Bipolar electrodes were placed to the left of the left eye and to the right of the right eye (HEOG) and above and below the right eye (VEOG) to record eye movement artefacts. Impedances were kept below 5 kΩ for the 64 recording electrodes and below 10 kΩ for the eye electrodes.

Before segmentation, continuous EEG activity was filtered low-pass using a zero phase shift digital filter with a cut-off frequency of 20 Hz and high-pass filter with a cut-off frequency of 0.1 Hz. After visual inspection to dismiss major artefacts (dismissing 157 trials out of 12240, i.e. 1.3 %), eye blinks were mathematically corrected based on the procedure advocated by Gratton, Coles and Donchin (1983). After correction, any trial with amplitude exceeding $\pm$ 100 μV at any point within an epoch and at any recording site except VEOG and HEOG was discarded from analysis (2.2 % of the remaining trials). Continuous EEG activity was segmented into epochs ranging from -100 to 1000 ms after stimulus onset. Baseline correction was performed in reference to pre-stimulus activity, and individual averages were digitally re-referenced to the global average reference. Only correct trials were kept for the analyses (dismissing 11.1 % of the remaining trials). Seven individual data sets were discarded due to excessive noise and/or alpha contamination resulting in undetectable early components (P1-N1 complex), leading to an average of 51.4 (SD = 5.6) trials per condition.

We proceeded with data analysis in two stages: (i) Comparing mean ERP amplitude differences between experimental conditions; (ii) Conducting a corrected Stroop-like conflict analysis with ERP data controlled for physical differences between conditions induced by *case-type*, i.e., after subtracting from each individual condition the signal contribution of *case-type* in order to validate the *congruence* effects.

(i) In the first analysis, we investigated the main effects of *word-type* and *case-type* and, critically, the interaction between the two factors. We measured the N1-P2 complex between 175–225 and 265–315 ms, respectively (that is 50 ms around grand-average peak times and in good agreement with component morphology in the literature), at posterior sites of predicted and observed maximal amplitude (P7, PO7, PO9, P8, PO8, and PO10).

The negative modulation by Stroop conflict was analysed between 350–450 ms (the predicted time-window based on previous studies, see e.g., Rebai, Bernard and Lannou, 1997; Liotti et al., 2000; West, 2003; Markela-Lerenc et al., 2004; Eppinger et al., 2007; Szűcs, Soltész and White, 2009; Szűcs and Soltész, 2010; Tillman and Wiens, 2011; Chen et al., 2011, the average N450 temporal window computed from these 9 studies is 354–460 ms) over 9 predicted fronto-centro-parietal electrodes (FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2).

In addition, we observed a late positive complex (LPC) between 500–700 ms at electrodes FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, and CP2 where it reached maximum observed amplitude.

(ii) In the corrected Stroop conflict analysis, we first subtracted the mean amplitude of CAPS and MiXEd combined averages across *all*, *some* and *case*, in each participant and each individual condition. In other words, this analysis compared the ERPs elicited by each of the critical stimuli after deducting Mean(ALL, SOME, and CASE) from CAPS conditions, and Mean(aLl, alL, ALl, aLL, SOme, SoMe, sOMe, soME, sOmE, cASe, CAse etc.) from MiXEd conditions.

We then analysed again the N1-P2 complex, N450 and LPC in the same temporal windows and at the same electrode locations as before.

In all cases, mean amplitudes, and peaks latency, were analysed using repeated measures ANOVAs with *case-type*, and *word-type* as within-subject factors.

2.3. Sentence-picture verification task

2.3.1. Materials

The sentence-picture verification task included 54 *some*-trials: 18 unambiguously true, that is, informative (see **Fig. 2A**), 18 unambiguously false (**Fig. 2B**) and 18 under-informative (**Fig. 2C**). Sentences were built on the following template: 'Some of the [circles/squares/triangles] are [blue/green/red]' and paired with an image depicting five coloured geometrical shapes. The 9 possible different associations of shapes and colours were presented twice but paired with different images. The sentence-picture verification task also comprised 342 filler trials involving *all*, *none*, *most*, *two*, conditional perfection (e.g., *Each of the shapes is green if it is a triangle*, see e.g., Geis and Zwicky, 1971), exhaustivity in *it*-clefts (e.g., *It is the squares that are blue*), and free choice inferences (e.g., *Each of the circles are red or blue*, see Chemla and Bott, 2014; van Tiel and Schaeken, 2016).

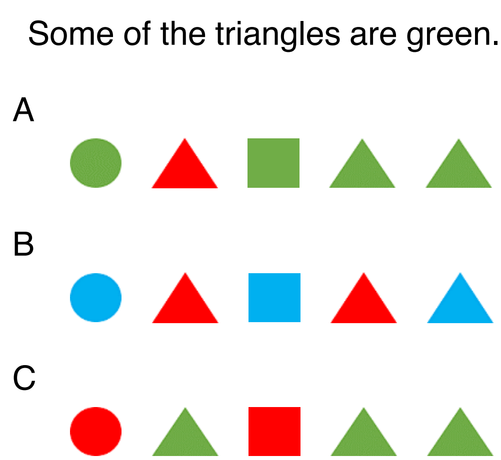

Figure 2: **Example of *some*-statement in the sentence-picture verification task. A.** Informative, **B.** False, and **C.** Under-informative trials.

2.3.2. Procedure

Participants were asked to indicate whether or not sentences were good descriptions pictures by pressing dedicated buttons on a response box. Response side was counterbalanced between participants, and the stimuli were randomly presented. When participants had read a sentence (presented for a maximum of 4 seconds), they could press any button for the associated picture to be displayed, and then had up to 3 seconds to provide their response. A short training session involving 14 unambiguous trials with feedback preceded the main task. Participants needed around 30 minutes to complete the task in 4 blocks, and were invited to take a break between blocks at their convenience.

3. Results

3.1. ERP experiment

3.1.1. Behavioural results

Accuracy (M = 88.2 %, SD = 32.25, see **Fig. 3A**) was analysed using logit mixed models[1] (see e.g., Jaeger, 2008) including the maximal random effect structure justified by the design (see e.g., Barr et al., 2013), namely by-participant random intercepts and by-participant random slopes for *word-type*, *case-type* and their interaction. We computed p-values by performing likelihood ratio tests in which a model with the fixed effect of interest was compared with an otherwise identical model without the said fixed effect.

The first model revealed a significant *word-type* $\times$ *case-type* interaction ($\chi^2(2) = 28.06$, $p < .001$). Analyses by *word-type* conditions showed that the interaction was driven by a significant effect of *case-type* restricted to *all* ($\chi^2(1) = 31.05$, $p < .001$; *some*: $\chi^2(1) = 1.97$, $p = .16$; *case*: $\chi^2(1) = 0.19$, $p > .6$). Analyses by *case-type* conditions showed no effect of *word-type* in the CAPS condition ($\chi^2(2) = 2.2$, $p > .3$), but a significant effect of *word-type* in the MiXEd

[1] Logit mixed models fitted using the R (R Core Team, 2014) package *lmerTest* (Kuznetsova, Bruun Brockhoff and Haubo Bojesen Christensen, 2014).

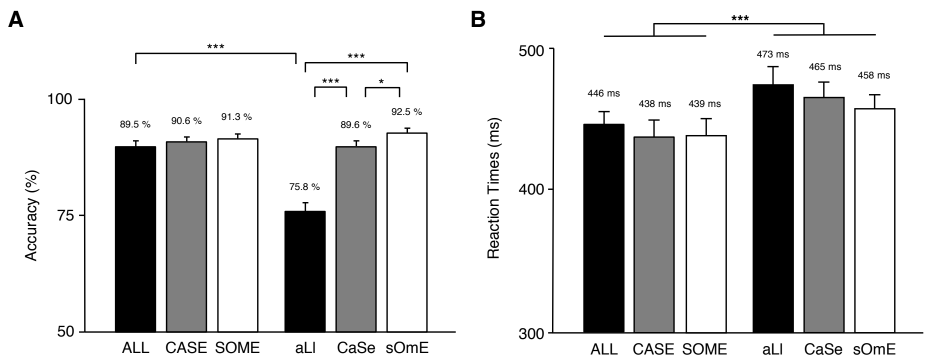

Figure 3: **Behavioural results. A.** Accuracy, **B.** Reaction times, for *all*, *case*, and *some* as a function of *case-type*. Integers indicate means, and error bars represent SEM by participants. *** $p < .001$, ** $p < .01$, * $p < .05$.

condition ($\chi^2(2) = 44.4$, $p < .001$). Tukey's post-hoc tests[2] within the MiXEd condition showed that accuracy was significantly higher for *some* relative to *case* ($\beta = 0.41$, $SE = 0.15$, $z = 2.69$, $p < .05$) and *all* ($\beta = 1.51$, $SE = 0.16$, $z = 9.48$, $p < .001$), and for *case* relative to *all* ($\beta = 1.1$, $SE = 0.15$, $z = 7.3$, $p < .001$).

Reaction times (see **Fig. 3B**) were analysed using linear mixed models[3] (see e.g, Bates, 2005; Baayen, Davidson and Bates, 2008; Baayen and Milin, 2010) including maximal random structure justified by the design, that is, by-participant random intercepts and by-participant random slopes for *case-type*, *word-type* and their interaction. The final models included removal of outliers (data points with absolute standardised residuals exceeding 2.5 standard deviations, see e.g, Baayen and Milin, 2010). Only reaction times from accurate responses were kept for the analysis, 11.8 % of data points were therefore dismissed, leaving 8574 data points out of 9720.

There was a significant effect of *case-type* on reaction times (F(1,26) = 23.7,

[2]Tukey's post-hoc tests were performed using the *glht* function of the R package *multcomp* (Hothorn, Bretz and Westfall, 2008).

[3]Linear mixed models fitted using the R package *lmerTest*.

$p < .001$, $\eta_p^2 = .46$)[4]. RTs to *all* and *case* increased by around 27 ms in the MiXEd compared with the CAPS condition, but only by around 19 ms for *some*, see **Fig. 3B**. However, there was no interaction between *case-type* and *word-type* ($p > .5$).

3.1.2. ERP results

Grand-average ERP waveforms are displayed in **Fig. 4**, **5**, and **6**. CAPS and MiXEd conditions markedly differed in ERP amplitude starting at around 180 ms after stimulus onset (see **Fig. 4**).

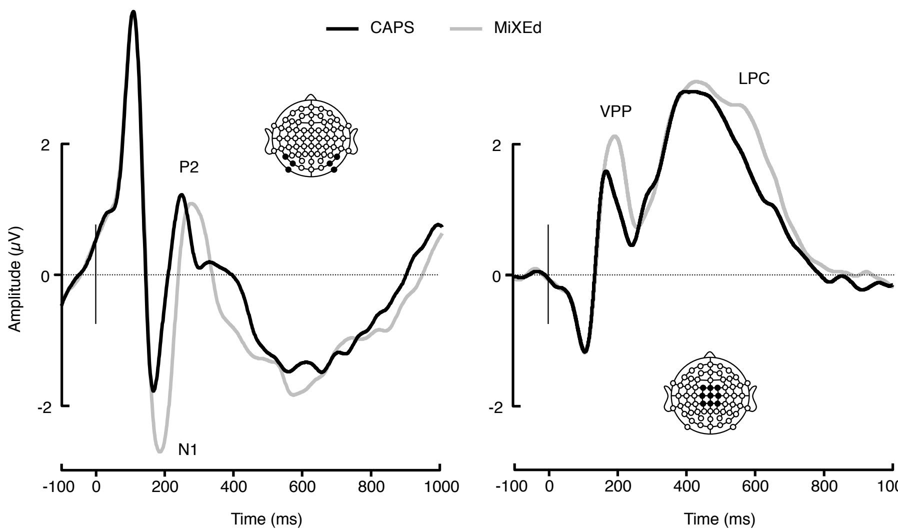

Figure 4: **ERP modulations elicited by *case-type*.** *Left*, Grand-average ERP waveforms elicited over the parieto-occipital region (linear derivation of P7, PO7, PO9, P8, PO8, PO10) in the CAPS (black line) and MiXEd (grey line) conditions; *Right*, Grand-average ERP waveforms elicited over the central region (linear derivation of FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2) in the CAPS (black line) and MiXEd (grey line) conditions.

N1 mean amplitudes were significantly modulated by *case-type* ($F(1,26) = 87.9$, $p < .001$, $\eta_p^2 = .77$), see **Fig. 4** and **5**. This factor marginally interacted with *word-type* ($F(2,52) = 2.7$, $p = .077$, $\eta_p^2 = .09$), due to the N1 effect varying

[4]We used the *anova* function of *lmerTest* which provides analysis of variance tables of type 3 with denominator degrees of freedom calculated based on Satterthwaite's approximation.

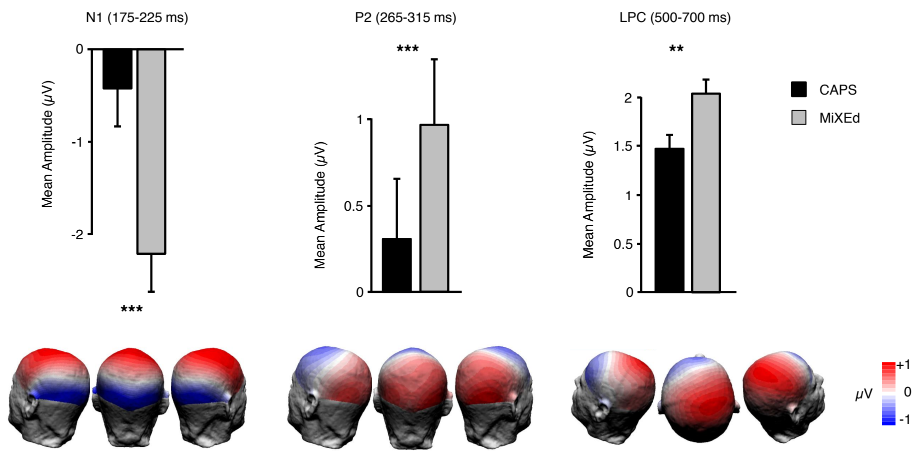

Figure 5: **N1, P2 and LPC mean amplitudes as a function of *case-type*, and MiXEd minus CAPS topographies.** *Left*, N1 (175–225 ms) mean amplitudes in the parietal-occipital region elicited in the CAPS and MiXEd conditions, and MiXEd minus CAPS N1 topography; *Middle*, P2 (265–315 ms) mean amplitudes in the parietal-occipital region elicited in the CAPS and MiXEd conditions, and MiXEd minus CAPS P2 topography; *Right*, LPC (500–700 ms) mean amplitudes in the central region elicited in the CAPS and MiXEd conditions (error bars represent SEM), and MiXEd minus CAPS LPC topography. *** $p < .001$, ** $p < .01$, * $p < .05$.

slightly in magnitude between word conditions (*all*: M = -1.41 μV; *case*: M = -1.95 μV; *some*: M = -2.08 μV). Nonetheless, the MiXEd condition elicited significantly greater amplitude than the CAPS condition for all word-types (*all*: F(1,26) = 32.1, $p < .001$, $\eta^2 = .55$; *case*: F(1,26) = 50.12, $p < .001$, $\eta^2 = .66$; *some*: F(1,26) = 61.3, $p < .001$, $\eta^2 = .70$). N1 mean peak latencies were marginally modulated by *case-type* (F(1,26) = 3.2, $p < .1$, $\eta^2 = .11$), in the absence of an interaction with *word-type* ($p > .6$). MiXEd elicited slightly delayed N1s compared to CAPS[5].

P2 mean amplitudes were also significantly modulated by *case-type* (F(1,26) = 14.32, $p < .001$, $\eta^2 = .35$), but this factor did not interact with *word-type* ($p > .5$). MiXEd elicited greater P2 amplitudes than CAPS words, see **Fig. 4** and **5**. P2 mean peak latencies were not significantly modulated by *case-type* (F(1,26) = 2.5, $p = .13$), and there was no interaction with *word-type* ($p > .6$).

Critically, in the N450 range, previously established as the Stroop conflict time-window, mean amplitudes were marginally modulated by *word-type* (F(2,52) = 2.45, $p < .1$, $\eta_p^2 = .09$), were not modulated by *case-type* ($p > .6$), and there was a significant interaction between the two factors (F(2,52) = 11.18, $p < .001$, $\eta_p^2 = .30$). Analyses by *word-type* conditions showed a significant effect of *case-type* for *all* (F(1,26) = 6.7, $p < .05$, $\eta^2 = .20$), and *some* (F(1,26) = 5.6, $p < .05$, $\eta^2 = .18$), but not for *case* (F(1,26) = 1.54, $p = .226$), see **Fig. 6**. MiXEd (incongruent) *all* significantly increased N450 mean amplitudes relative to CAPS (congruent) *all*. As for *some*, this effect was reversed, that is, N450 mean amplitudes were significantly increased for CAPS relative to MiXEd. There was no significant difference between CAPS and MiXEd *case*.

LPC mean amplitudes, see **Fig. 4** and **5**, were significantly modulated by *case-type* (F(1,26) = 12.3, $p < .01$, $\eta^2 = .32$), such that MiXEd elicited greater LPC amplitudes than CAPS, but this factor did not interact with *word-type* ($p > .6$).

[5]Note that the N1 effects reported here were mirrored over central regions of the scalp in the form of a vertex positive potential (VPP, see e.g, Eimer, 2011), see **fig. 4** and **6**.

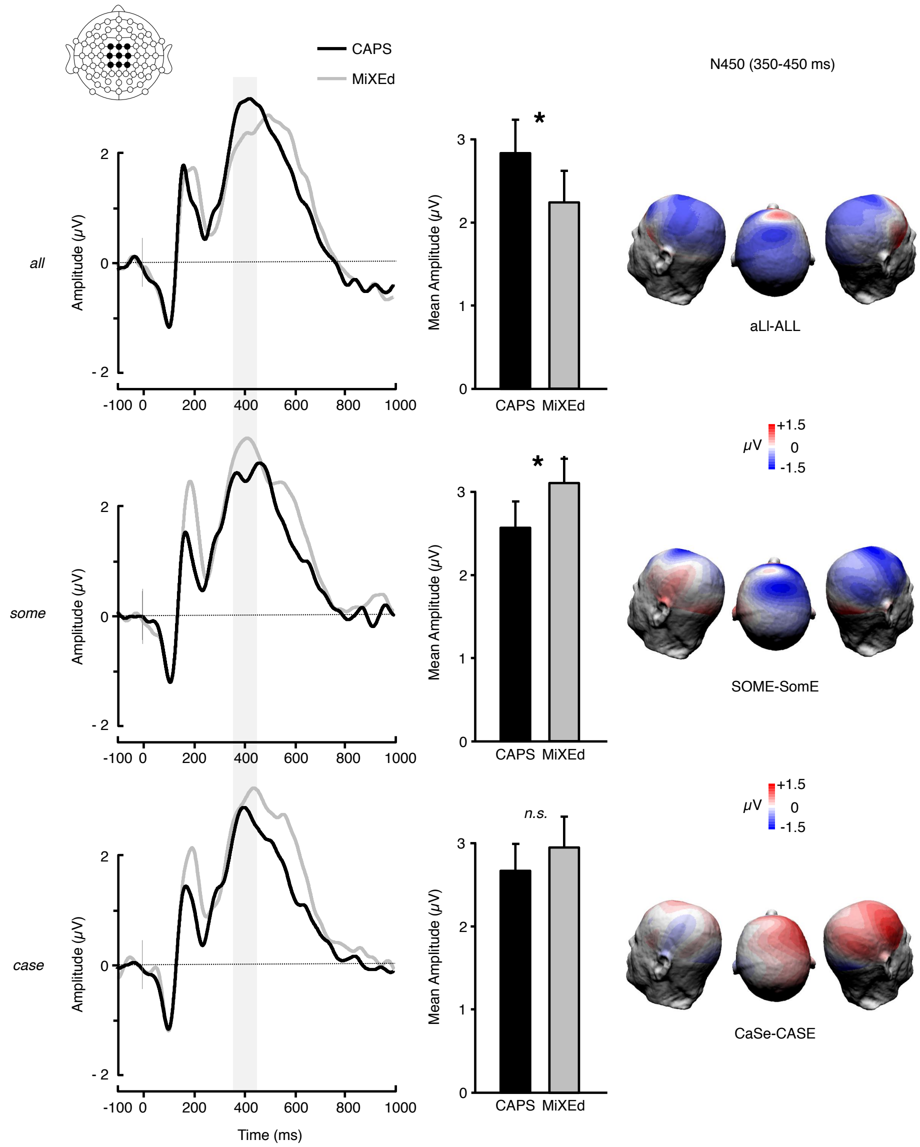

Figure 6: **Effect of *case-type* and *congruence* in the N450 range.** Grand-average ERP waveforms elicited over the central region (linear derivation of FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2) in the CAPS (solid black line) and MiXEd (solid grey line) conditions, mean amplitudes in the central region between 350–450 ms as a function of *case-type*, and N450 effect (incongruent minus congruent) topography, for *all* (top), *some* (middle), and *case* (bottom). *** $p < .001$, ** $p < .01$, * $p < .05$.

We then proceeded to analysing ERP amplitudes after correction of physical differences (see Methods section 2.2.3). There was no main effect of *case-type* on N1 mean amplitudes ($p > .1$), but there was a marginal interaction between *case-type* and *word-type* (F(2,52) = 2.7, p = .08, η_p^2 = .08) driven by stimulus *all* (F(1,26) = 6.2, $p < .05$, η^2 = .19). There was no main effect of *case-type* or interaction between *case-type* and *word-type* on N1 mean peak latency of corrected ERPs, but a marginal main effect of *word-type* (F(2,52) = 2.9, p = .07) driven by a significant difference between *some* and *case* (Bonferroni adjusted $p < .05$).

Corrected P2 mean amplitudes were not significantly modulated by either *case-type* or *word-type*, and there was no interaction between the two factors (all $ps > .1$). Corrected P2 latencies were only affected by *word-type* (F(2,52) = 3.8, $p < .05$), an effect driven by a significant difference between *case* and *some* (Bonferroni adjusted $p < .05$).

In the N450 range, the effect of *case-type* was preserved (F(1,26) = 9.8, $p < .01$, η_p^2 = .27) and the interaction between *case-type* and *word-type* was significant (F(2,52) = 11.2, $p < .001$, η_p^2 = .30). Analyses by *word-type* showed that the effect of *case-type* was significant for the semantic test condition *all* (F(1,26) = 17.2, $p < .001$, η^2 = .40), the pragmatic test condition *some* (F(1,26) = 10.2, $p < .01$, η^2 = .28), but not the neutral control condition *case* (F(1,26) = 1.538, p = .226), see **Fig. 7**. Furthermore, when excluding the neutral condition *case* from the analysis, and after conversion of the *case-type* factor to *congruence* since the meaning of ALL was congruent with its physical form whereas that of SOME was pragmatically incongruent, the N450 effect proved of lesser magnitude for *some* than *all*, as shown by a marginal *congruence* × *word-type* interaction (F(1,26) = 3.95, p = .057, η_p^2 = .13, see **Fig. 7**).

Finally, LPC effects did not survive physical differences correction (all $ps >$ .5).

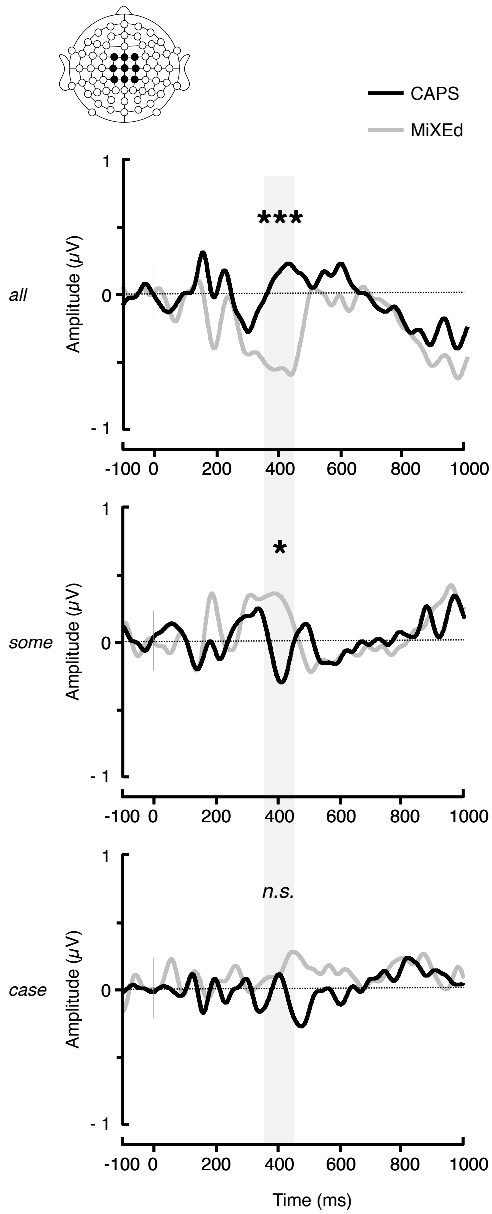

Figure 7: **Stroop-like conflict effect on ERPs corrected for physical differences.** Grand-average corrected ERP waveforms elicited over the central region (linear derivation of FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2) in the CAPS (solid black line) and MiXEd (solid grey line) conditions for *all* (top), *some* (middle), and *case* (bottom). *** $p < .001$, ** $p < .01$, * $p < .05$.

3.2. Sentence-picture verification task

Logical response rates were high overall (Mean = 0.79, Median = 0.94, SD = 0.28; see **Fig. 8**). Six participants gave 9 or less logical responses out of 18, 12 participants gave 10 to 17 logical responses, and 9 participants only gave logical responses. Therefore, only 6 participants out of 27 could be considered “pragmatic”.

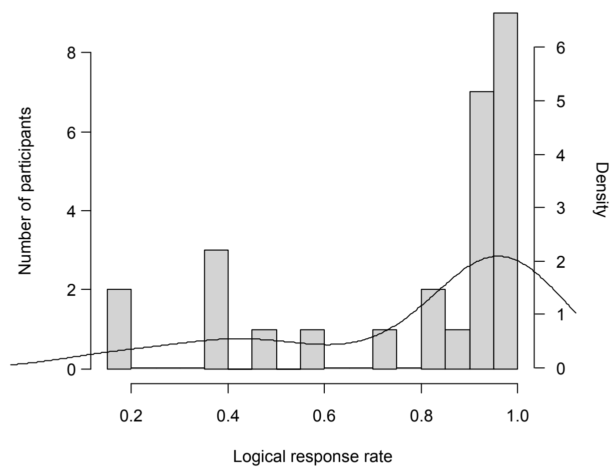

Figure 8: **Distribution of logical responses to under-informative *some*-statements.**

There was no significant correlation between the number of logical responses and the N450 effect elicited by *some* in the ERP experiment (i.e. incongruent SOME minus congruent e.g., SoMe N450 amplitude, r = -.24, p = .22).

4. Discussion

Using ERPs in a Stroop-like paradigm involving upper case and mixed case letter strings, we investigated whether the scalar inference ‘not all’ triggered by *some* is derived when *some* is presented in isolation, i.e. in a situation of minimal context.

As expected, longer RTs for MiXEd than CAPS stimuli showed that the MiXEd condition proved more demanding than the CAPS condition (see e.g., Mayall, Humphreys and Olson, 1997; Mayall, Humphreys, Mechelli et al., 2001;

Juhasz et al., 2006; Arditi and Cho, 2007; Lien, Allen and Crawford, 2012). Significant differences in accuracy expected from the Stroop conflict only appeared for *all*. ERPs collected simultaneously also revealed a main effect of physical form in N1, P2 and LPC time-windows, with greater amplitudes and/or delayed latencies generally observed for MiXEd as compared to CAPS stimuli. Critically, raw ERPs, and ERPs corrected for physical differences, showed significant increases in amplitude for incongruent relative to congruent stimuli in the predicted N450 time-window. *Some*, in particular, produced the ERP modulations that one would expect if its meaning were construed pragmatically.

Accuracy results obtained in the *all* condition suggest that our Stroop conflict manipulation worked: Performance was hindered for incongruent MiXEd *all* as compared to congruent CAPS *all*. We note however that, in the case of incongruent *all*, 'l' having the same physical height as 'L' as in e.g., 'ALl', could have relatively increased the magnitude of the Stroop-like conflict for *all*[6]. Furthermore, the relative lack of a clear behavioural Stroop effect in the case of *some* (e.g., when comparing SOME vs. SomE) could be due to the MiXEd condition (in which *some* was congruent under our hypothesis) being overall more demanding as discussed above. Nevertheless, accuracy in this condition was higher for *some* than for *case*, suggesting that *some* was processed with more ease than *case* when its pragmatic meaning ('some but not all') was consistent with its guise. The absence of a clear behavioural Stroop effect for *some* may also be due to its suboptimal appearance (from a pragmatic point of view) when presented in MiXEd form. Indeed, a stimulus featuring two upper case and two lower case letters is compatible with quantifiers such as 'half' or 'two'. Whereas upper case *all* may be considered perfectly congruent, and mixed case *all*, perfectly incongruent with the meaning of *all*, *some* only enjoys a perfect status in its upper case, pragmatically incongruent form (SOME). In its pragmatically congruent form (e.g., sOmE), half or two rather than merely some of

[6]After removal of the ambiguous occurrences of MiXEd *all* ('ALl' and 'AlL') from the analysis, the Stroop-like effect on accuracy vanished (mean accuracy incongruent *all* = 92 %).

the letters are in upper case, which is pragmatically suboptimal, because *some* is taken to mean at least/more than one and less than half, and numbers tend to be preferred to *some* in the subitising range (see e.g., Kaufman et al., 1949; Degen and Tanenhaus, 2015). Considering the fact that the word has only four letters, we could not implement a pragmatically optimal form of *some* in this experiment.

The overall processing difficulty of the MiXEd relative to the CAPS condition was also reflected in event-related brain potentials through greater mean amplitudes in the N1, P2 and LPC ranges[7]. The modulation by *case-type* in the N1 and P2 ranges reflects difficulty with visual feature-to-letter mapping as predicted by models of visual word recognition (see e.g., Bi-modal interactive activation model (BIAM), Grainger and Holcomb, 2009). The N1 modulation observed for MiXEd as compared to CAPS is compatible with the kind of modulation observed previously when letters are rotated in words (see Kim and Straková, 2012). The interaction between *case-type* and *word-type* further suggested a reduced *case-type* effect for *all*, probably because the letter 'l' has the same height as an upper case letter (whereas all letters in *some* or *case* have a different height in upper and lower case), or because 'l' can be confused with 'I', making its status as a lower case letter ambiguous. After correction for physical differences, a marginal *congruence* × *word-type* interaction remained, driven by *all*, arguably due to the correction process: The *case-type* modulation was smaller in the *all* than in the *some* and the *case word-type* conditions, resulting in a harsher correction for *all* than for *some* or *case*, thus artificially increasing the CAPS/congruent vs. MiXEd/incongruent difference. Note that the *case-type* effect which extended into the LPC window in the raw ERP analysis, disappeared

[7]Note that we found no sign of a P1 amplitude modulation by low-level physical differences between *case-type* conditions. This is surprising given the controversy in the face-processing literature regarding the sensitivity of the P1 to differences between stimuli in terms of e.g., size, disparity and shape (see Thierry et al., 2007; Dering et al., 2011 vs. Rossion and Jacques, 2008; Eimer, 2011).

after correction for physical differences.

The Stroop conflict revealed by an N450 modulation (see e.g., Liotti et al., 2000; West, 2003; Markela-Lerenc et al., 2004; Szűcs, Soltész and White, 2009; Szűcs and Soltész, 2010; Tillman and Wiens, 2011) was obtained in the semantic test condition *all* (e.g., aLl vs. ALL), as well as the pragmatic test condition *some* (SOME vs. e.g., SoMe), indicating that *some* presented in isolation is construed as incongruent with 'all', that is, even when extraction of its meaning is irrelevant and not required to complete the task at hand. The Stroop conflict effect elicited by *some* was however less strong than that elicited by *all*. This needs not indicate that, albeit pragmatic in origin, the meaning of *some* was less compelling than that of *all*. Indeed, this may be interpreted as a task/stimulus effect: As noted above, whereas *all* may be considered perfect in its congruent (ALL) or incongruent (e.g., aLL) guises, *some* only enjoys a perfect status in its pragmatically incongruent guise (SOME).

The N450 conflict effect observed for SOME is overall incompatible with a strong context-dependency view of the SI 'not all', given that in a situation of minimal linguistic context, SOME is not construed as congruent with the concept of 'all'. Our results suggest that the 'not all' meaning of *some* might be represented in memory rather than computed on the fly as suggested by some theoretical accounts (see e.g., Sperber and Wilson, 1995; Carston, 2004), and thus that *some* may be ambiguous in the mental lexicon. Psycholinguistic studies have provided empirical evidence that multiple meanings of ambiguous words are activated during early stages of processing, before a single interpretation is ultimately selected based on meaning frequency and context (see e.g., Swinney, 1979; Onifer and Swinney, 1981). However, lexical and semantic information retrieval is expected to occur at around 200 ms post-stimulus onset (see e.g., Martin et al., 2014; Hauk et al., 2012), and thus the N450 effect reported here cannot inform lexical-semantic stages of information processing. Overall, our results are consistent with the view that the SI is the predicted meaning of *some* in the absence of special circumstances (Levinson, 2000, see Introduction section 1). Our results may also be consistent with a grammatical account according to which SIs are derived within grammar and are not pragmatic in essence. Indeed, it seems that a specific context is needed for SIs not to arise according to such view: A silent operator *Only* is inserted whenever possible, and SIs are not licensed only in specific contexts (see e.g., Chierchia, 2004; Chierchia, Fox and Spector, 2012).

Our results show that *some* does not need a strong contextual support to appear inconsistent with the meaning 'all'. An alternative hypothesis could be that *all* appearing as frequently as *some*, might have rendered the lexical scale $\langle$*all*, *some*$\rangle$ (see e.g., de Carvalho et al., 2016) and thus the contrast between *some* and *all* salient to participants despite the presence of a high number of filler and distractor trials. More likely, making all/not all CAPS decisions may have elicited an implicit context calling for an 'all in CAPS?' Question Under Discussion (see Introduction section 1). Such QUD may have resulted in word association participating in the Stroop conflict. Indeed, the words *some* and *all* are strongly associated (see e.g., Edinburgh Associative Thesaurus, Kiss et al., 1973), and psycholinguistic studies have shown that interference effects are modulated by association strength: Colour words that are less strongly associated with the concept of colour produce less Stroop interference (see e.g. Scheibe, Shaver and Carrier, 1967; Proctor, 1978). Therefore, the Stroop conflict found for SOME could also partly originate from the association between the words *some* and *all*. However, the 'all in CAPS?' QUD fails to account entirely for the effect observed for *some* because it predicts a Stroop-like effect also for *case* since neither SOME nor CASE are forms of the word *all*. Furthermore, even if an 'all in CAPS?' QUD in the experimental context was sufficient to support the SI 'not all', it remains that the SI was irrelevant in the task at hand, and the results therefore do not support the view according to which SI derivation is dependent upon contextual relevance.

A strong context-driven model of SI, or at least of the 'not all' SI triggered by *some* (generalising to all SIs would probably be flawed, see e.g., van Tiel, van Miltenburg et al., 2014), according to which an SI is only computed when it is contextually relevant, is not supported by our data. Reading the word *some* appears sufficient to evoke inconsistency with ‘all’. Note that we are not arguing here for context-insensitivity: Even if the SI is evoked by *some* in isolation, hearers should be more or less committed to it depending on contextual information.

Interestingly, most of the participants of this study could be considered “logical” since almost 80 % of the under-informative *some*-statements were considered good descriptions in the sentence-picture verification task. Yet, the same participants exhibited a Stroop-like conflict when presented with the pragmatically incongruent stimulus SOME in the ERP experiment. This seems to indicate that “logical” behaviour may stem from cognitive strategising rather than mere linguistic processing. At least, this result calls for caution when making hypotheses concerning processing on the basis of metalinguistic judgements.

5. Conclusion

In this study, we aimed at testing the context-dependency of the SI ‘not all’ triggered by *some*. In order to do so, we presented the quantifier in a minimal, that is, single word context in a Stroop-like task. An N450 modulation, marker of Stroop conflict, elicited by the pragmatic incongruent CAPS *some*, similar to that found for the semantic incongruent MiXEd *all*, revealed that the quantifier on its own can evoke the inference ‘not all’, or at least can be considered inconsistent with ‘all’. This argues for context-*sensitivity* rather than strong context-*dependency* of SI derivation in the case of *some*: The scalar inference ‘not all’ should be construed as more or less salient and relevant depending on the context rather than entirely contingent upon it. Further research is however required to clarify interactions between lexical-semantics, default heuristics, predicted meaning, and genuine pragmatic processing.

Funding

This work was supported by the Swiss National Science Foundation (grant numbers P2NEP1_155426 and P300P1_164558, CB).

Acknowledgements

The authors thank Si Jing Tan, Emma Jones, Yang Li, Ondřej Zíka, and Jiehui Hu for their assistance with data collection.

References

Arditi, Aries and Jianna Cho (2007). "Letter case and text legibility in normal and low vision". In: *Vision research* 47.19, pp. 2499–2505. DOI: 10.1016/j.visres.2007.06.010.

Baayen, R. Harald, D. J. Davidson and Douglas Bates (2008). "Mixed-effects modeling with crossed random effects for subjects and items". In: *Journal of Memory and Language* 59, pp. 390–412.

Baayen, R. Harald and Petar Milin (2010). "Analyzing reaction times". In: *International Journal of Psychological Research* 3.2, pp. 12–28.

Barr, Dale J, Roger Levy, Christoph Scheepers and Harry J Tily (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal". In: *Journal of memory and language* 68.3, pp. 255–278.

Bates, Douglas (2005). "Fitting linear mixed models in R". In: *R news* 5.1, pp. 27–30.

Beaver, D and B Clark (2008). *Sense and Sensitivity: How focus determines meaning (Explorations in Semantics)*. Oxford: Blackwell.

Bergen, Leon and D. J. Grodner (2012). "Speaker knowledge influences the comprehension of pragmatic inferences". In: *Journal of Experimental Psychology: Learning, Memory & Cognition* 38 (5), pp. 1450–1460.

Bonnefon, Jean-François, Aidan Feeney and Gaëlle Villejoubert (2009). "When some is actually all: Scalar inferences in face-threatening contexts". In: *Cognition* 112.2, pp. 249–258.

Bott, Lewis and Ira Noveck (2004). "Some utterances are underinformative: the onset and time course of scalar inferences". In: *Journal of Memory and Language* 51, pp. 437–457.

Breheny, Richard, Heather Ferguson and Napoleon Katsos (2013). “Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation”. In: *Language and Cognitive Processes* 28.4, pp. 443–467.

Breheny, Richard, Napoleon Katsos and John Williams (2006). “Are generalised scalar implicatures generated by default?” In: *Cognition* 100 (3), pp. 434–463.

Carston, Robyn (2004). “Truth-conditional content and conversational implicature”. In: *The semantics / pragmatics distinction*, pp. 65–100.

Chemla, Emmanuel and Lewis Bott (2014). “Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice”. In: *Cognition* 130.3, pp. 380–396. DOI: http://dx.doi.org/10.1016/j.cognition.2013.11.013.

Chen, Antao, Kira Bailey, Brandy N Tiernan and Robert West (2011). “Neural correlates of stimulus and response interference in a 2–1 mapping stroop task”. In: *International Journal of Psychophysiology* 80.2, pp. 129–138.

Chevallier, Coralie, Ira Noveck, Lewis Bott, V. Lanza, T. Nazir and Dan Sperber (2008). “Making disjunctions exclusive”. In: *Quarterly Journal of Experimental Psychology* 61 (11), pp. 1741–1760.

Chierchia, Gennaro (2004). “Scalar implicatures, polarity phenomena and the syntax/pragmatic interface”. In: *Structures and Beyond*. Ed. by A. Beletti. Oxford: Oxford University Press, pp. 39–103.

Chierchia, Gennaro, Danny Fox and Benjamin Spector (2012). “Scalar implicature as a grammatical phenomenon”. In: 3. Ed. by Maienborn, von Heusinger and Portner, pp. 2297–2331.

de Carvalho, Alex, Anne C Reboul, Jean-Baptiste Van der Henst, Anne Cheylus and Tatjana Nazir (2016). “Scalar Implicatures: The psychological reality of scales”. In: *Frontiers in psychology* 7.

De Neys, Wim and Walter Schaeken (2007). “When People Are More Logical Under Cognitive Load. Dual Task Impact on Scalar Implicature”. In: *Experimental Psychology* 54 (2), pp. 128–133.

Degen, Judith and Michael K Tanenhaus (2015). “Processing Scalar Implicature: A Constraint-Based Approach”. In: *Cognitive science* 39.4, pp. 667–710.

Dering, Benjamin, Clara D Martin, Sancho Moro, Alan J Pegna and Guillaume Thierry (2011). “Face-sensitive processes one hundred milliseconds after picture onset”. In: *Frontiers in human neuroscience* 5, p. 93.

Eimer, Martin (2011). “The face-sensitive N170 component of the event-related brain potential”. In: *The Oxford handbook of face perception* 28, pp. 329–44.

Eppinger, Ben, Jutta Kray, Axel Mecklinger and Oliver John (2007). “Age differences in task switching and response monitoring: Evidence from ERPs”. In: *Biological Psychology* 75.1, pp. 52–67.

Geis, M.L. and A.M. Zwicky (1971). “On invited inferences”. In: *Linguistic inquiry* 2.4, pp. 561–566.

Geurts, Bart (2010). *Quantity implicatures.* Cambridge University Press.

Goodman, Noah D and Andreas Stuhlmüller (2013). “Knowledge and implicature: Modeling language understanding as social cognition”. In: *Topics in cognitive science* 5.1, pp. 173–184.

Grainger, Jonathan and Phillip J Holcomb (2009). “Watching the word go by: On the time-course of component processes in visual word recognition”. In: *Language and linguistics compass* 3.1, pp. 128–156. DOI: 10.1111/j.1749-818X.2008.00121.x.

Gratton, Gabriele, Michael GH Coles and Emanuel Donchin (1983). “A new method for off-line removal of ocular artifact”. In: *Electroencephalography and clinical neurophysiology* 55.4, pp. 468–484.

Grice, H. P. (1975). “Logic and conversation”. In: *Syntax and Semantics*, pp. 41–58.

Grodner, D. J., N. M. Klein, K. M. Carbary and M. K. Tanenhaus (2010). ““Some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment”. In: *Cognition* 116.1, pp. 42–55.

Hartshorne, Joshua K, Jesse Snedeker, Stephanie Yen-Mun Liem Azar and Albert E Kim (2015). “The neural computation of scalar implicature”. In: *Language, cognition and neuroscience* 30.5, pp. 620–634.

Hauk, O, C Coutout, A Holden and Y Chen (2012). “The time-course of single-word reading: evidence from fast behavioral and brain responses”. In: *Neuroimage* 60.2, pp. 1462–1477.

Horn, L. R. (1972). “On the Semantic Properties of Logical Operators in English”. PhD thesis. UCLA.

— (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

Hothorn, Torsten, Frank Bretz and Peter Westfall (2008). “Simultaneous Inference in General Parametric Models”. In: *Biometrical Journal* 50.3, pp. 346–363.

Huang, Yi Ting and Jesse Snedeker (2009). “Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface”. In: *Cognitive Psychology* 58.3, pp. 376–415.

Hunt, Lamar, Stephen Politzer-Ahles, Linzi Gibson, Utako Minai and Robert Fiorentino (2013). “Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification”. In: *Neuroscience Letters* 534, pp. 246–251.

Jaeger, T.F. (2008). “Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models”. In: *Journal of Memory and Language* 59, pp. 434–446.

Juhasz, Barbara J., Simon P. Liversedge, Sarah J. White and Keith Rayner (2006). “Binocular coordination of the eyes during reading: Word frequency and case alternation affect fixation duration but not fixation disparity”. In: *The Quarterly Journal of Experimental Psychology* 59.9, pp. 1614–1625. DOI: 10.1080/17470210500497722.

Kaufman, Edna L, Miles W Lord, Thomas Whelan Reese and John Volkmann (1949). “The discrimination of visual number”. In: *The American journal of psychology* 62.4, pp. 498–525.

Kim, Albert E and Jana Straková (2012). “Concurrent effects of lexical status and letter-rotation during early stage visual word recognition: evidence from ERPs”. In: *Brain research* 1468, pp. 52–62.

Kiss, George R, Christine Armstrong, Robert Milroy and James Piper (1973). "An associative thesaurus of English and its computer analysis". In: *The computer and literary studies*, pp. 153–165. URL: http://www.eat.rl.ac.uk/.

Kuznetsova, Alexandra, Per Bruun Brockhoff and Rune Haubo Bojesen Christensen (2014). *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-20. URL: http://CRAN.R-project.org/package=lmerTest.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge: MIT Press.

Lien, Mei-Ching, Philip A Allen and Caitlin Crawford (2012). "Electrophysiological evidence of different loci for case-mixing and word frequency effects in visual word recognition". In: *Psychonomic bulletin & review* 19.4, pp. 677–684. DOI: 10.3758/s13423-012-0251-9.

Liotti, Mario, Marty G Woldorff, Ricardo Perez and Helen S Mayberg (2000). "An ERP study of the temporal course of the Stroop color-word interference effect". In: *Neuropsychologia* 38.5, pp. 701–711.

MacLeod, Colin M (1991). "Half a century of research on the Stroop effect: an integrative review." In: *Psychological bulletin* 109.2, p. 163.

Markela-Lerenc, Jaana, Nicole Ille, Stefan Kaiser, Peter Fiedler, Christoph Mundt and Matthias Weisbrod (2004). "Prefrontal-cingulate activation during executive control: which comes first?" In: *Cognitive Brain Research* 18.3, pp. 278–287.

Martin, Clara, X. Garcia, A. Breton, Guillaume Thierry and Albert Costa (2014). "From literal meaning to veracity in two hundred milliseconds". In: *Frontiers in Human Neuroscience* 8, pp. 1–12.

Marty, Paul P and Emmanuel Chemla (2013). "Scalar implicatures: working memory and a comparison with only". In: *Frontiers in Psychology* 4.403. DOI: 10.3389/fpsyg.2013.00403.

Mayall, Kate, Glyn W Humphreys, Andrea Mechelli, Andrew Olson and Cathy J Price (2001). "The effects of case mixing on word recognition: Evidence from a PET study". In: *Journal of Cognitive Neuroscience* 13.6, pp. 844–853. DOI: 10.1162/08989290152541494.

Mayall, Kate, Glyn W Humphreys and Andrew Olson (1997). "Disruption to word or letter processing? The origins of case-mixing effects." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23.5, p. 1275. DOI: 10.1037/0278-7393.23.5.1275.

Nieuwland, Mante, Tali Ditman and Gina Kuperberg (2010). "On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities". In: *Journal of memory and language* 63.3, pp. 324–346.

Noveck, Ira and A. Posada (2003). "Characterising the time course of an implicature". In: *Brain & Language* 85, pp. 203–210.

Onifer, W. and D. A. Swinney (1981). "Accessing lexical ambiguities during sentence comprehension: effects of frequency of meaning and contextual biais". In: *Memory and Cognition* 15, pp. 225–236.

Politzer-Ahles, Stephen and Robert Fiorentino (2013). "The Realization of Scalar Inferences: Context Sensitivity without Processing Cost". In: *PloS one* 8.5, e63943.

Politzer-Ahles, Stephen, Robert Fiorentino, Jiang Xiaoming and Zhou Xiaolin (2013). "Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification". In: *Brain Research* 1490, pp. 134–152.

Politzer-Ahles, Stephen and Laura Gwilliams (2015). "Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography". In: *Language, Cognition and Neuroscience* 30.7, pp. 853–866.

Proctor, Robert W (1978). "Sources of color-word interference in the Stroop color-naming task". In: *Perception & Psychophysics* 23.5, pp. 413–419. DOI: 10.3758/BF03204145.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. version 3.1.0. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rebai, Mohamed, Christian Bernard and Jacques Lannou (1997). “The Stroop’s test evokes a negative brain potential, the N400”. In: *International Journal of Neuroscience* 91.1-2, pp. 85–94.

Roberts, Craige (1996). “Information structure in discourse: Towards an integrated formal theory of pragmatics”. In: *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pp. 91–136.

Rossion, Bruno and Corentin Jacques (2008). “Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? Ten lessons on the N170”. In: *Neuroimage* 39.4, pp. 1959–1979.

Scheibe, Karl E, Phillip R Shaver and Samuel C Carrier (1967). “Color association values and response interference on variants of the Stroop test”. In: *Acta psychologica* 26, pp. 286–295. DOI: 10.1016/0001-6918(67)90028-5.

Sperber, Dan and Deirdre Wilson (1995). *Relevance. Communication and cognition*. 2nd ed. Oxford: Blackwell. 340 pp.

Swinney, D. A. (1979). “Lexical access during sentence comprehension: (Re)consideration of context effects”. In: *Journal of Verbal Learning and Verbal Behaviour* 18, pp. 645–659.

Szűcs, Dénes and Fruzsina Soltész (2010). “Stimulus and response conflict in the color–word Stroop task: a combined electro-myography and event-related potential study”. In: *Brain research* 1325, pp. 63–76.

Szűcs, Dénes, Fruzsina Soltész and Sonia White (2009). “Motor conflict in Stroop tasks: Direct evidence from single-trial electro-myography and electro-encephalography”. In: *Neuroimage* 47.4, pp. 1960–1973.

Thierry, Guillaume, Clara D Martin, Paul Downing and Alan J Pegna (2007). “Controlling for interstimulus perceptual variance abolishes N170 face selectivity”. In: *Nature neuroscience* 10.4, pp. 505–511.

Tian, Ye, Heather Ferguson and Richard Breheny (2016). “Processing negation without context–why and when we represent the positive argument”. In: *Language, Cognition and Neuroscience* 31.5, pp. 683–698.

Tillman, Carin M and Stefan Wiens (2011). “Behavioral and ERP indices of response conflict in Stroop and flanker tasks”. In: *Psychophysiology* 48.10, pp. 1405–1411.

van Tiel, Bob and Walter Schaeken (2016). “Processing Conversational Implicatures: Alternatives and Counterfactual Reasoning”. In: *Cognitive Science*, n/a–n/a. ISSN: 1551-6709. DOI: 10.1111/cogs.12362. URL: http://dx.doi.org/10.1111/cogs.12362.

van Tiel, Bob, Emiel van Miltenburg, Natalia Zevakhina and Bart Geurts (2014). “Scalar diversity”. In: *Journal of Semantics*, pp. 1–39. DOI: 10.1093/jos/ffu017.

West, Robert (2003). “Neural correlates of cognitive control and conflict detection in the Stroop and digit-location tasks”. In: *Neuropsychologia* 41.8, pp. 1122–1135.